埼玉女子短期大学研究紀要　第4号　1993.03

類似係数とクラスター化法

Similarity and clustering method

高林茂樹
Shigeki Takabayashi

This paper offers a survey of similarity and clustering method. It also shows cluster analysis about questionary on student's studying computer. This cluster analysis found that students consist of three clusters.

1．はじめに

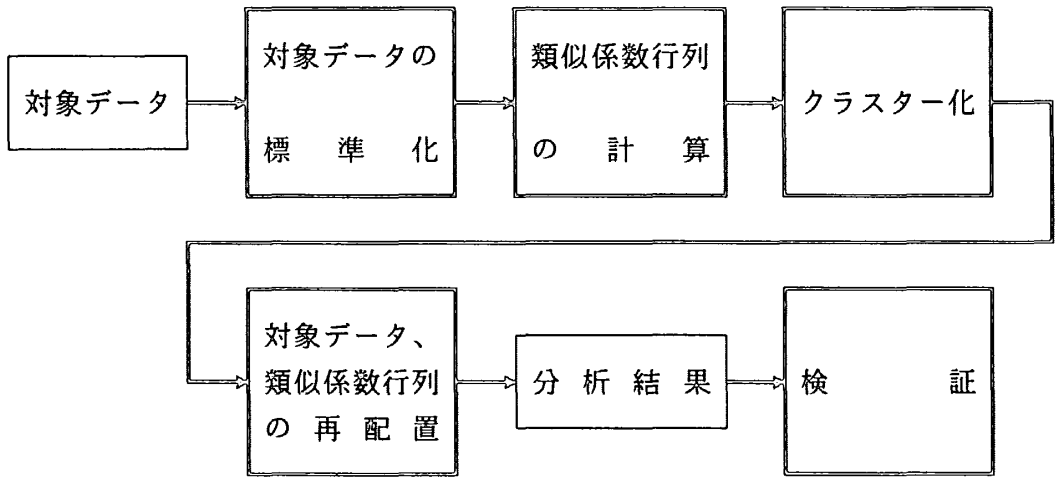
対象となるものの中から、類似しているものを見つけ出し数学的に分類する場合に用いられるのが、クラスター分析である。因子分析や主成分分析、数量化理論Ⅲ類、数量化理論Ⅳ類などで分析すると対象データが空間座標上に位置づけられることが分かり、いくつかのグループに分類されそうなことも推測できる。異なるグループに属すると思われるメンバー間で距離が離れていて境界が明瞭な場合は、特別な判定を行わなくともグループに分けることは容易であるが、境界が明瞭でなかったり、空間が3次元以上では、グループつまりクラスター分けを視覚的に行うことは難しい。このような時、クラスター分析を用いて、対象となるデータのメンバー間の類似度つまり類似係数を計算し、類似しているメンバーをまとめて行くことにより分類することができる。

クラスター分析では、類似係数の計算方法とそれを使ってのクラスター化の方法に各種のものがある。したがって研究目的や対象データの性質などによってこれらを使い分ける必要がある。本論文は、類似係数とクラスター化に対して考察を加えたあと、コンピュータの授業に関しての学生に対するアンケートをもとに因子分析した数値を用いて、クラスター分析を行い、学生層の分類を試みたものである。

2．クラスター分析の方法

クラスター分析を行う場合、一般的には、まず対象データを標準化し、類似係数行列を計算しクラスター化する。その結果を樹形図で表し、対象データと類似係数行列を見やすいように樹形図に合わせて再配置する。最後に類似係数行列のクラスター化前とクラス

ター化後の相関を計算するなどして、結果の検証をする。



図－1　クラスター分析のフロー

標準化は、データの持つ属性間の大きさをそろえるために行ない下記により計算される。

$$\frac{Xi - m}{s}$$

Ｘｉ：対象データ　　ｍ：平均　　ｓ：標準偏差

2－1．類似係数の計算

類似係数は、データの性質によりさまざまなものが考えられている。大別すると、量的なデータに対するものと質的なデータに対するものに分けることができる。

（1）量的なデータに対する類似係数

X_{ik}でｉはｉ番目の対象、ｋはｋ番目の属性を表す。ｎは属性の個数とすると量的なデータに対する類似係数では、次のものが使われている。

◦ユークリッドの距離係数

$$e_{ij} = \{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2\}^{1/2}$$

・平均ユークリッドの距離係数

$$d_{ij}=\{\sum_{k=1}^{n}(X_{ik}-X_{jk})^{2}/n\}^{1/2}$$

・形状差係数

$$Z_{ij}=\{n/(n-1)\ (d_{ij}^{2}-q_{ij}^{2})\}^{1/2}$$

$$q_{ij}^{2}=(1/n^{2})\ (\sum_{k=1}^{n}X_{ik}-\sum_{k=1}^{n}X_{jk})^{2}$$

・コサイン係数

$$C_{ij}=\frac{\sum_{k=1}^{n}X_{ik}X_{jk}}{(\sum_{k=1}^{n}X_{ik}^{2})^{1/2}\ (\sum_{k=1}^{n}X_{jk}^{2})^{1/2}}$$

・相関係数

$$r_{ij}=\frac{\sum_{k=1}^{n}X_{ik}X_{jk}-\sum_{k=1}^{n}X_{ik}\sum_{k=1}^{n}X_{jk}/n}{[\{\sum_{k=1}^{n}X_{ik}^{2}-(\sum_{k=1}^{n}X_{ik})^{2}/n\}\ \{\sum_{k=1}^{n}X_{jk}^{2}-(\sum_{k=1}^{n}X_{jk})^{2}/n\}]^{1/2}}$$

・キャンベラ距離係数

$$a_{ij}=(1/n)\sum_{k=1}^{n}\frac{|X_{ik}-X_{jk}|}{(X_{ik}+X_{jk})}$$

・ブレイ・カーティス係数

$$b_{ij}=\frac{\sum_{k=1}^{n}|X_{ik}-X_{jk}|}{\sum_{k=1}^{n}(X_{ik}+X_{jk})}$$

これらのうちコサイン係数と相関係数は、－1と1の範囲にあり類似度が高いほど大きな値となる。それ以外は、類似度が高いほど小さな値となり非類似係数と言われるものである。ユークリッドの距離係数、平均ユークリッドの距離係数、形状差係数は、0と∞の範囲にあり、キャンベラ距離係数、ブレイ・カーティス係数は0と1の範囲にある。

（2）質的なデータに対する類似係数

データが有／無や色など数量として表せない場合に使われる係数である。

X_{ik}でiはi番目の対象、kはk番目の属性を表す。nは属性の個数とし、属性は2値で、0または1の場合を見ることにする。X_{ik}とX_{jk}のn個の属性の中で1－1のものがa個、1－0のものがb個、0－1のものがc個、0－0のものがd個とするとよく使われる質的なデータに対する類似係数C_{ij}には、次のものがある。

◦Ｊａｃｃａｒｄ係数

$$C_{ij}=\frac{a}{a+b+c}$$

◦単純見合い係数

$$C_{ij}=\frac{a+d}{a+b+c+d}$$

◦Ｓｏｒｅｎｓｏｎ係数

$$C_{ij}=\frac{2a}{2a+b+c}$$

これらの係数は、いずれも0と1の範囲にあり、類似度が高いほど大きな値となる。

2－2．クラスター化の方法

クラスター化とは、類似係数行列を基にして、類似したクラスターをまとめてクラスター数を減らすことである。まとめる段階で、類似係数行列を計算し直す方法により主な方法には次のものがある。

（1）ＵＰＧＭＡ(unweighted pair - group method using arithmetic averages)

クラスターｐとｑがまとまって１つのクラスターになった場合、それ以外のクラスターｒと今度新たにまとまったクラスターとの係数をｒとｐ、ｒとｑの間の類似係数の平均として計算する方法である。

（2）ＳＬＩＮＫ(single linkage clustering method)

クラスターｐとｑがまとまって１つのクラスターになった場合、それ以外のクラスターｒと今度新たにまとまったクラスターとの係数をｒとｐ、ｒとｑの間の類似係数の中で大きい方の値とする方法である。非類似係数の場合は小さい方の値となる。

（3）ＣＬＩＮＫ(complete linkage clustering method)

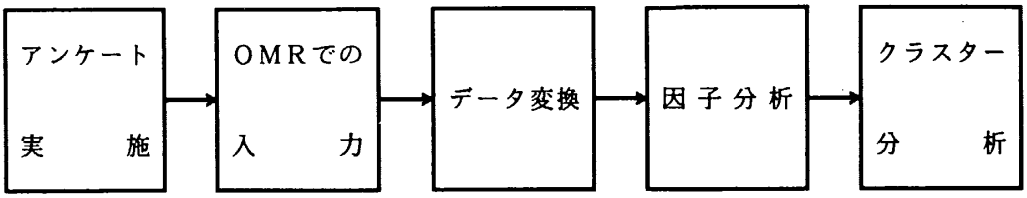
クラスターｐとｑがまとまって１つのクラスターになった場合、それ以外のクラスターｒと今度新たにまとまったクラスターとの係数をｒとｐ、ｒとｑの間の類似係数の中で小さい方の値とする方法である。非類似係数の場合は大きい方の値となる。

（4）Ｗａｒｄ法

平方和指標Ｅをそれぞれのまとめ方ごとに計算し、Ｅの値が最小になる場合を選んでまとめていく方法である。Ｅは、クラスターの仮の集合について類似係数の平均を計算し、計算前の各類似係数との差を求めて平方し合計したものである。

3．クラスター分析の例

コンピュータの授業についての学生アンケートをＯＭＲ(Optical Mark Reader)を用いて読み取り、データ変換をしたあと因子分析を行い、そこで計算された因子得点をもとにクラスター分析を行った。なお、因子分析とクラスター分析の計算には、㈱社会情報サービスの「統計解析シリーズ」を使用した。



図－２　アンケート実施からクラスター分析までのフロー

3－1．アンケートの実施

アンケートは、コンピュータの授業に関して、無記名で設問は8種類66項目について207名の学生に実施した。内容は、「説明、準備等に対する満足度」、「授業方法などでの良い点」、「授業方法などでの悪い点」、「学習効果」、「学生の授業態度での良い点」、「学生の授業態度での悪い点」、「空き時間の利用、検定等の勉強」、「コンピュータでやりたいこと」についてである。この中の前半は、「学生の声　調査表」による設問を利用した。

3－2．OMRでの入力

アンケートの結果は、すべてOMR用のシートに記入し、OMRで読み取り、フロッピーディスクに記録した。読み取りにあたっては1コラムに複数回答できるようにプログラミングした。

3－3．データ変換

回答中には、量的データが11個、質的データが55個あったので、計算誤差や計算時間、計算結果の解釈などを考慮して8種類の量的データに変換した。質的データを量的データに変換する方法は、複数回答の中のデータを分類してシートにマークされた個数を量的データとした。さらにそれらのデータを「統計解析シリーズ」の入力フォーマットに変換した。

3－4．因子分析

因子分析では、3つの因子を抽出し、バリマックス回転後の因子負荷量は、表－1のとおりである。因子1は、「空き時間の利用、検定等の勉強」、「コンピュータでやりたいこと」に、因子2は、「説明、準備等に対する満足度」、「授業方法などでの良い点」、「樹業方法などでの悪い点」に、因子3は、「学習効果」、「学生の授業態度での悪い点」に関して相関あるいは逆相関が見られる。

表－1　因子分析による因子負荷量（バリマックス回転後）

質　　問	因子　1	因子　2	因子　3
説明、準備等に対する満足度	0.060920	0.675237	－0.054791
授業方法などでの良い点	0.255694	0.716772	0.208365
授業方法などでの悪い点	0.144162	－0.713105	0.255000
学習効果	0.204279	0.198349	－0.675087
学生の授業態度での良い点	0.329670	0.444409	0.150578
学生の授業態度での悪い点	0.037037	0.079988	0.825057
空き時間の利用、検定等の勉強	0.851001	0.034915	－0.016185
コンピュータでやりたいこと	0.863824	0.064516	－0.049304

3－5．クラスター分析

クラスター分析は、類似係数または非類似係数としてユークリッドの距離係数、クラスター化はＣＬＩＮＫを用いた。クラスターが8個の状態でのそれぞれのデータ数が表－2である。また各クラスターと因子分析で求めた因子負荷との関係を、平均、標準偏差、最小値、最大値で見ると表－3のようになる。

表－2　クラスター別データ数

クラスター	データ数
1	29
2	16
3	41
4	22
5	15
6	29
7	38
8	17
合　計	207

表－3　因子別クラスター統計値

因子1

クラスター	平　　均	標準偏差	最　小　値	最　大　値
1	−0.17737	0.10595	−0.37776	0.01028
2	0.12435	0.11485	0.00310	0.42157
3	0.19391	0.14809	0.00060	0.65063
4	−0.12437	0.08095	−0.32592	0.01434
5	0.24081	0.17112	0.00500	0.53644
6	0.16667	0.12165	0.02784	0.51026
7	−0.22247	0.11514	−0.47266	0.03671
8	−0.12069	0.09293	−0.39088	0.02136

因子2

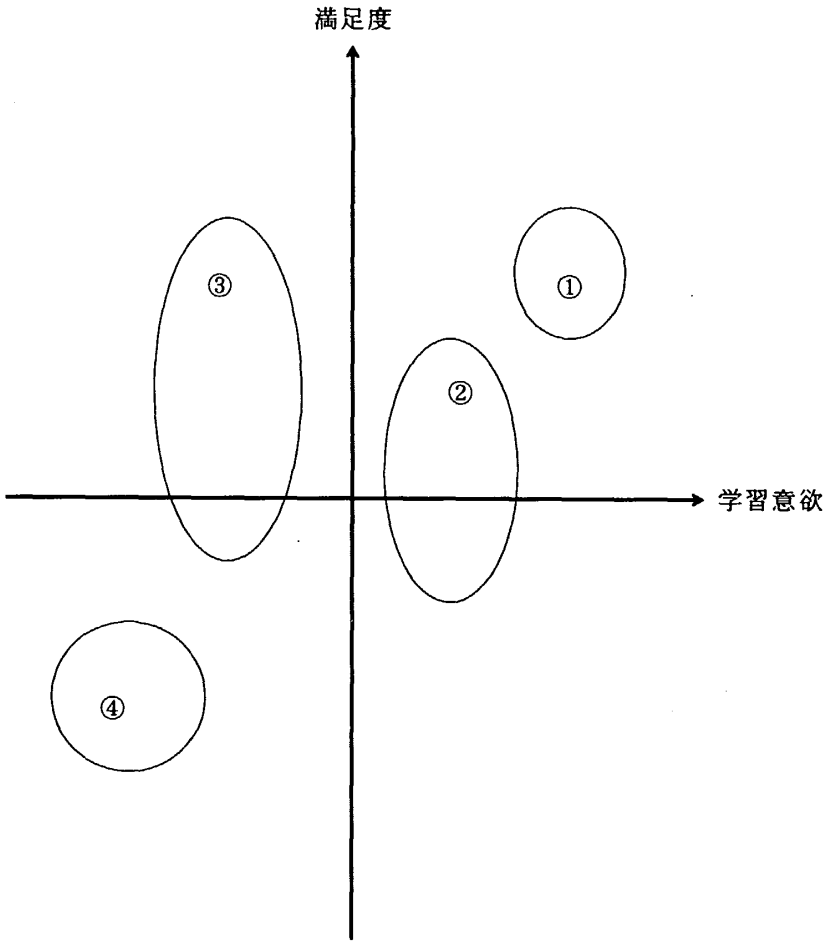
クラスター	平　　均	標準偏差	最　小　値	最　大　値
1	−0.15606	0.09693	−0.50693	−0.00746
2	−0.12998	0.10141	−0.38647	−0.00489
3	0.16204	0.11056	0.00420	0.50653
4	0.12969	0.09603	0.00352	0.34446
5	−0.15585	0.16700	−0.57210	−0.00697
6	0.21747	0.14767	0.00668	0.52312
7	−0.23706	0.18077	−0.73476	−0.00599
8	0.12637	0.08541	0.00058	0.27781

因子3

クラスター	平　　均	標準偏差	最　小　値	最　大　値
1	−0.10742	0.10167	−0.35138	−0.00170
2	−0.08492	0.07184	−0.23358	−0.00362
3	−0.13143	0.07353	−0.26998	−0.00420
4	−0.10127	0.09301	−0.33128	−0.00420
5	0.13384	0.15689	0.00306	0.53242
6	0.13744	0.14086	0.00237	0.75342
7	0.11330	0.09369	0.01229	0.46182
8	0.10541	0.13034	0.00284	0.51480

3－6．結果のまとめ

クラスター分析で求めた8個のクラスターを表－1、表－2、表－3をもとに学習意欲、授業満足度、学習態度または学習効果の観点から整理すると図－3のようになる。図の①～④は次のとおりである。



図－3　クラスター分析の結果

① 学習意欲もあり授業に対する満足度も高いグループ。（20%）

② 学習意欲はあるグループ。（29%）

このなかで不満はないが学習効果がない。（14%）

不満がある。（15%）

③ 学習意欲がないグループ。（33%）

このなかで満足していないが学習効果はある。（14%）

満足している。（11%）

④ 学習意欲もなく授業に対して不満もあるグループ。（18%）

この結果は、因子分析後の因子負荷量を用いているため標準化されたデータで計算したものである。したがって、満足度も学習意欲も相対的なものであって絶対的なものではな

い。しかし、この結果で見る限りにおいては、大きく分けて、2：6：2の割合で、学習意欲も高く授業にも満足していると思われるグループ、学習意欲・満足度は普通のグループ、学習意欲もあまりなく満足もしていないグループになると考えられる。

4．おわりに

クラスター分析は、生物学、社会科学、自然科学、考古学などいろいろな分野で使われている。しかしながら類似係数の計算やクラスター化の方法が多数あるため研究目的や対象データの性質などをよくつかんで対応するなど工夫が必要となる。クラスター化の方法を変えて計算して結果が大幅に異なる場合には、クラスターがはっきり別れていない可能性がある。このような場合は樹形図を吟味することになるが、データ件数が多いと樹形図全体を吟味するのはかなり困難となる。統計的な処理全体に言えることであるが、計算されたものが真理ではないにもかかわらず、真理とされてしまう危険はできるだけ避けなければならない。そのためには、方法を変えたりデータを増やしたりして、さらに計算することが必要である。

（参考文献）

（1）H.C.Romesberg、西田英郎、佐藤嗣二訳、「実例クラスター分析」、内田老鶴圃、1992

（2）管民郎、「多変量解析」、社会情報サービス、1991